基于语料库和规则库的人名识别法*

北京工业大学计算机学院人工智能研究室　　中国人民大学中国语言文学系
宋　柔　　朱　宏　　潘维桂　　尹振海

摘要：人名识别是分词系统实用化必须解决的问题。我们将基于计算机统计的语料库方法和基于人工归纳的规则库方法结合起来，对新闻文章的人名加以识别，其准确度可达９５％以上，而且速度相当快。

Automatic Recognition of Person Names based on Copus and Rule_base

Rou Song　Hong Zhu　　Weigui Pan　Zhenhai Yin

Artificial Intelligence Lab.　　Chinese Language Department
Beijing Computer Institue　　Renmin University of China
Beijing 100044, China　　Beijing 100872, China

Abstract: The recognition of person names is an important problem in chinese word segmentation system ,especially in practical system.This paper presents a method based on copus and rule_base, which finds out more than 95% of person names with high speed.

1. 引言

分词是高层次汉语处理的基础。现在各种分词系统和分词算法已出现很多，准确率可达９５％以上，但在面对日常语料时，便显示出在一些关键问题上还很不成熟，在实用化方面还需要做大量的工作。这些关键问题包括：

歧义字段归并
人名识别
地名识别
企业字号识别
商标牌名识别
缩略语识别

我们正在新闻语料库上工作，首当其冲的是遇到大量名不见经传的人名、地名、字号和商标名。我们探索了一些方法，在人名处理方面已取得较为满意的结果。

关于人名识别，已发表的成果不多，我们见到的只有台湾清华大学张俊盛等的文章【1】。该项工作的原理是对姓名语料库和一般语料库作统计分析，通过动态规划，选择词汇出现机率之乘积最高者作为分词结果，同时也识别出了人名。我们的作法是将基于计算机统计的语料库方法和基于人工归纳的规则库方法结合起来。

2. 人名语料库及其统计处理

实际语料中常遇到的人名大体可分为三类：汉人名及类汉人名（如朝鲜人名、越南人名等），西方人名和日本人名。我们从目前的实用需要出发，主要研究大陆汉人名，也注意了英美人的大陆用汉译名。

关于英美人的汉译名用字，有《世界译名手册》可参考，但该书条目太庞杂。我们收录的是《新英汉词典》附录《常见英美姓名表》中的绝大多数字，共４００余个。这个姓名表虽然不是来自实际的英美人名语料库，而且只有２４００余人名，但它代表了大陆近几十年形成的欧美人汉译名用字规范，故很有代表性。我们的策略是：凡是由这些单字（用最大匹配法不能聚合成词的字）组成的字串，中间可有一个或两个圆点，便是待定的英美人名。从已检验过的例子看，这样选出的待定人名涵盖了几乎所有西方人（包括俄罗斯人）的汉译名。

关于大陆汉人名，我们选取了北京某城区１２万人的名字作为姓名语料库。由于该地区居民以国家机关工作人员及其家属为主，他们来自全国各地的城市和农村，故该语料库有较好的代表性。我们从中排除了“张王氏”一类特殊构造的人名后，在两级国标字的范围内分别统计了姓氏字、单名字、双名首字、双名末字的用字频率。统计结果显示：

⑴ 姓氏字的前几十个高频字，其频率确有区别意义。比如这１２万人使用了８３６个姓，其中前３姓（王、张、李）覆盖２５%，前１３姓覆盖５０%，前４００姓覆盖９９%。但较低频字的频率，相互之间并无区别意义。这方面，中国科学院遗传所以大陆人口１／２０００的抽样作姓氏统计，他们也有相似的结论。

⑵ 名字用字频率的区别意义更差。如这１２万人用的单名字共１６６８个，前3字（军、伟、静）只覆盖４%，覆盖９９%需１３１３字，占总数的７９%。

⑶ 单名字和双名首字的用字范围是不同的，主要是某些双名首字中的高频字在单名字中是甚低频字，甚至不用作单名字。如双名首字的最高频字“淑”在该语料库中作为单名字仅１次，作为双名首字排第８的“小”不作单名字出现。

⑷ 双名首字和末字之间有相关性。如“小”作为双名首字排第８，“珍”和“生”作为双名末字分别排第５和第８，但该１２万人中没有叫“小珍”或“小生”的，“爱人”、“富农”等也是类似情况。

根据统计和分析的结果，我们对这四类字采取了如下的处理方法：

⑴ 对于姓氏字，大致取前４００字，覆盖率为９９%。对其中的低频姓氏字作少许人工调整，以减少偶然性带来的误差。

⑵ 对其它三类字，在语料库中出现３次和３次以上的都取。只出现１次或２次的字中，属于二级国标字的原则上都取，属于一级国标字的原则上不取。原因是二级国标字使用频率很低，在现代汉语新闻语料中除了用于姓名外，很少用于其它。一级国标字的活性很大，收得太宽易造成人名误判。

⑶ 凡收取的字，只区别它属于这四类中的哪一类或哪几类，不计算它在类中的频率或排名。

⑷ 属于双名首字类的字和双名末字类的字，可一前一后组成相当多的双字词。对这些词用人工进行筛选，区分出能用作双名的词和一般不用作双名的词。

在处理实际语料时首先按词典进行分词，然后以姓氏字和复姓词为标记找出候定的汉人名。其检测条件是：

⑴ 姓氏字（词）后跟一个可用作双名的双字词。

⑵ 姓氏字（词）后跟一个双名首字类字，再跟一个双名末字类字。

⑶ 姓氏字（词）后跟一个单名字类字。

3. 人名识别规则

人名语料库的统计结果为找出待定人名提供了依据。将待定人名确认为人名，我们依据的不是统计而是规则。规则的原理是：在日常见到的语料，尤其是新闻语料中，首次提到一个不见经传的人名时，一般要在其人名前或后加一些限制成分，作为读者对这个陌生人认知的出发点。以后再提到此人时，便可不加限制成分而只提其名。比照程序设计语言的术语，我们把这第一次出现称为该名的定义性出现，以后的出现为该名的使用性出现。

依据这个原理，我们每处理一篇文章都造二个人名表--分为汉人名表（包括类汉人名）和英美人名表。找出一个待定人名后，先看它是否在人名表中。如果在，它就是人名的使用性出现，可以确认为人名；如果不在，但带有适当的限制性成分或满足有关规律，它就是人名的定义性出现，可确认为人名，并加入人名表，否则不能确认为人名。

这里的所谓限制成分，第一类是用来表示人的身份的，我们称它们为身份词，包括表示职务、职业、头衔的词语和亲属称谓词语等。其中有些只能用在人名前面，称为前身份词，如“工人”、“教师”、“影星”、“犯人”、“丈夫”、“妻子”等；有些只用在人名后面，称为后身份词，如“同志”、“女士”等；还有些既可用在人名前，又可用在人名后，即兼属两类，如“教授”、“总理”、“小姐”等。“先生”、“太太”、“奶奶”等也是兼类，只是用在人名前后意义不同。

表示职衔的身份词前有时加修饰词，如“副总理”的“副”，“代部长”的“代”。当用作后身份词时，一般只加单字修饰词，如可以说“张百发副市长”和“常务副市长张百发”，不说“张百发常务副市长”。

表示身份的词语是难以列举完的，也没必要列举完，因为许多这类词语带有易识别的后缀字，如“在逃犯”、“理发员”、“面包师”、“目击者”中的“犯”、“员”、“师”、“者”。这类词语一般只用作前身份词。

第二类限制成分是地名和单位名，它们用在人名前面，表明该人所在地或所在单位。如“静海县大丘庄禹作敏”，“中国足球队古广明”。

第三类限制成分是较复杂的定语，加在人名前面，与人名之间通常用“的”相连。如“刚毕业的赵晓华”，“年过七旬的王贵芝”。

有少数情形是人名的定义性出现在句首，人名后面是有关的说明。如“马丁·路德·金是著名黑人律师”。

与汉人名相当的词语有“刘大娘”、“李四小姐”、“小张”、“老杨”、“小A”、“老Z”等，它们无需前后文的说明即能被确认。

成串人名之间用顿号和“和”、“与”、“同”、“及”等连词连结，此时限制性成分放在名字串的头或尾。放在名字串尾时，与最后一个人名之间加“等”，如“聂卫平、马晓春、刘小光等围棋国手”。

4. 计算机实现和运行效果

我们建立了一个近6万词的词库，并建立了一个为分词（包括人名、地名识别）服务的词类标记系统。人名语料库的统计分析结果和规则所需的信息大部分都吸收在这个标记系统中。分词过程是先做机械分词，使用的是从左向右的最大匹配算法。然后做歧义字段处理，最后依据规则处理数词、西文、简单动词短语，单位名、地名和人名等。程序用MSC600编制，可在IBM PC及其兼容机上运行。

我们从《经济日报》印刷厂随机选取了20篇已录入的文章，从中摘取所有含人名较多的段落，组成约1万字的试验样本。该语料在OCTEC 386/33上被处理，除去读入词库、读入语料、写出分词结果的时间，CPU运行时间4秒，平均每秒处理2500汉字。

该语料中有人名114个次，被正确识别108个次。有4个未能识别，都是歧义字段未能正确处理

而造成的。如“实习生肖弦义”，“实习生”未收在词库中，但“实习”与“生肖”都在词库中，从而被分成“实习　生肖　弦义”。有2个是误判，都与“向”有关。一个是“汽车司机向陈国良学习”，把“向陈国”误判为人名。另一个是“向王军这位普普通通的民警表示感谢”，把“向王军”误判为人名。此外有一个多判：“刘从陈衣袋里抢走现金４００元”，把“刘从陈”判为人名。依照【1】的标准，人名识别召回率是９４．７%，准确率是９７．４%。如改进歧义字段处理方法，再对一些特殊的姓氏字作特殊处理，效果估计会更好。

5. 讨论

我们认为，语料库方法可以排除人对语言规律认识中的主观随意性，并且在人们对纷繁复杂的语言现象无力驾驭时，语料库方法确能在一定范围内提供有效的帮助。比如，若不对人名语料库作统计分析，光凭人的感觉来决定人名用字，效果一定不佳。但这种方法存在严重的缺陷。根本问题在于自然语言的各层次，特别是短语层至篇章层，至今尚无比较令人满意的统计模型，甚至是否存在这种模型还很成问题。因此，各取所长，也许能合理有效地解决问题。

参考文献

【1】张俊盛等：多语料库作法之中文姓名辨识，《中文信息学报》，１９９２，３
【2】张国煊等：汉语自动分词的直接匹配算法及其词典结构，《机器翻译研究进展》，电子工业出版社，１９９２
【3】黄昌宁：语料库语言学，《中国计算机用户》，１９９０，１１
【4】刘开瑛等：自然语言处理，科学出版社，１９９０

---

* 本研究由国家自然科学基金及北京市自然科学基金资助

附录：　试验样本分词结果：

／／今年／／２／／月／／２５／／日／／上午／／１１／／时／／许／／，／／河北省／／秦皇岛市／／海港区／／文化路／／地区／／，／／警笛／／长／／鸣／／，／／一／／辆／／辆／／消防车／／风驰电掣／／般／／地／／开／／往／／海港／／区／／八三／／里／／生活／／小区／／。／／只／／见／／一／／栋／／居民／／楼／／浓／／烟／／滚滚／／，／／大火／／冲天／／。／／当／／消防／／队员／／冲／／进／／着火／／的／／室内／／时／／，／／发现／／一／／名／／少妇／／仰面／／躺／／在／／北／／屋／／地下／／，／／此／／人／／已／／奄奄一息／／，／／送／／到／／医院／／这／／名／／少妇／／已经／／死亡／／。／／海港／／公安分局／／接／／到／／报案／／后／／，／／局长／／王广新／／、／／副／／局长／／郭向甫／／迅速／／组织／／干警／／赶／／到／／现场／／。／／同时／／，／／秦皇岛／／市委／／常委／／、／／市／／公安局／／局长／／刘金国／／、／／副／／局长／／刘荣华／／带领／／技／／侦／／人员／／也／／赶／／赴／／现场／／指导／／破案／／。／／现场／／位于／／海港／／区／／八三／／里／／生活／／小区／／１３／／栋／／１０／／号／／室内／／，／／两／／室／／一／／厅／／，／／屋／／内／／家具／／、／／电器／／、／／衣物／／等／／全部／／被／／火烧毁／／。／／技术员／／经过／／精心／／勘查／／，／／找／／到／／了／／三／／个／／起火／／点／／。／／法医／／检查／／，／／死／／者／／张华／／，／／后脑勺／／部／／有／／一／／小／／圆形／／钝／／器／／伤／／，／／脖子／／处／／有／／出血／／点／／，／／象／／是／／掐／／痕／／，／／死亡／／原因／／是／／一氧化碳／／中毒／／。／／案情分析／／会／／在／／海港／／公安分局／／小／／会议／／室／／进行／／了／／整整／／一／／个／／晚上／／。／／在／／他杀／／还是／／自杀／／的／／问题／／上／／，／／多数／／人／／认为／／是／／他杀／／，／／依据／／是／／：／／１．／／现场／／上／／找／／到／／三／／个／／起火／／点／／，／／说明／／不是／／自然／／着火／／，／／而是／／纵火／／；／／２．／／死／／者／／虽／／是／／一氧化碳／／中毒死亡／／，／／但／／从／／死／／者／／后脑勺／／的／／钝／／器／／伤／／和／／脖子／／的／／出血／／点／／看／／，／／死／／者／／是／／被／／物体／／击／／昏／／或／／掐／／昏／／后／／，／／一氧化碳／／中毒死亡／／的／／；／／３．／／从／／烧／／剩／／的／／物品／／看／／，／／象／／有／／搬／／动／／痕迹／／。／／市／／公安局／／和／／海港／／分局／／的／／领导／／果断／／决定／／立案／／侦查／／。／／市／／公安局／／和／／海港／／分局／／迅速成立／／了／／“／／２／／·／／２５／／”／／专案／／组／／。／／他们／／兵／／分／／三／／路／／，／／调查访问／／，／／扩大／／线索／／，／／最后／／得出／／的／／结论／／是／／：／／熟人／／作案／／。／／其／／依据／／有／／二／／：／／１／／、／／八三／／里／／生活／／小区／／是／／公寓／／式／／管理／／，／／平时／／一般／／生人／／进不去／／，／／门卫／／看管／／的／／非常／／紧／／，／／案／／发／／前／／据／／门卫／／介绍／／没有／／生人／／出入／／；／／２／／、／／死／／者／／生前／／

比较//谨慎//，//防范//意识//强//。//据//她//丈夫//高海平//介绍//，//有//一次//邮局//的//同志//到//他//家//安装//电话//，//因//张华//不//认识//这个人//被//拒//之//门//外//，//说//等//她//丈夫//回来//再//安//，//这//就//大大//缩小//了//侦查//范围//。//专案//组//决定//以//熟悉//死//者//的//４０//多//个//厂矿//企事业单位//为//摸//排//范围//，//以//１８//－//４５//岁//的//职工//干部//为重点//对象//，//展开//了//深入//细致//的//调查//。//四//天//过后//，//专案//组//从//数//百//名//与//死//者//熟悉//的//人中//筛选//出//９//个//重点//嫌疑//人员//。//再//进一步//核//查//，//９//人中//又//排除了//８//个//，//剩下//一//人//的//疑//点//越来越//大//。//他//叫//李国毅//，//男//，//现//年//４０//岁//，//与//死//者//同//住//一//幢//楼房//。//案//发//的//当天上午//九//点//，//有//人证明//李国毅//回//家//一次//。//专案//组//找//他//了解//情况//时//，//李//承认//９//点//回//家//修//自家//厕所//阀门//，//１０//点//到//体育场//门//前//的//体育//商店//找//石永胜//老师//联系//钢材//的//事//。//可//专案//组//找//李//的//妻子//谢芝如//了解//李//２５//日//上午//的//活动//时//，//谢//说//２５//日//上午//，//李//找//她//两//次//，//见面//就//问//人//来//了//吗//？//谢//问//：//“//什么//人//来//了//？//”//李//说//：//“//从//东北//搞//了//两//吨//石油//。//”//可//李//根本//没//提//到//石油//的//事//，//两//人//说话//有//矛盾//点//。//专案//组//顺//线//往//下//追//，//找//到//石永胜//老师//再次//核实//李//的//情况//，//石//说//：//“//２５//日//上午//李//根本//没//提//倒卖//钢材//的//事//。//”//石//还//提供//一//条//重要//线索//，//当时//李//拿//出//１０００//元//全//是//１０//元//面额//的//号码//连//着//的//新//钱//，//叫//石//兑换//１００//元//一//张//票面的//钱//，//石//给//兑换//了//８//张//，//现//门市//上//还//剩//５６//张//１０//元//票面的//人民币//。//２７//日//李//又//到//门市部//，//告诉//石老师//换钱//的//事//另//当//外人//说//。//据//调查//死//者//张华//家//正巧//存//有//１０００//元//十//元//面额//号码//连//着//的//人民币//。//这些//情况//表明//，//李国毅//有//重大//嫌疑//。//经过//市//局//、//分局//领导//同意//，//专案//组//依法//传讯//了//李国毅//。//在//有力//的//证据//面前//，//李国毅//终于//难于//自圆其说//精神//防线//彻底//崩溃//，//彻底//供认//了//全部//作案//过程//。//２//月//２５//日//早//８//时//，//李//去//单位//上班//，//因//单位//没//活//，//他//去//了//一//趟//妻子//的//单位//。//９//点//左右//李//返回//单位//借//了//两//个//管//钳子//和//一//个//１．５//磅//的//锤子//回//自家//修//厕所//下水道//。//２０//分钟//后//从//自家//出来//，//遇//上//从//市场//买//菜//回来//的//邻居//张华//。//张华//请//李//帮助//修//张//家//的//厕所//下水管道//，//李//便//随//张//到家//，//把//厕所//下水管道//修好//后//，//见//张//的//丈夫//不在家//，//顿生//邪念//。//心想//，//早//听说//张//家//有钱//，//何不//借//此//机会//干掉//她//弄//点//钱//。//于是//，//李//将//张华//骗//进//厕所//，//拿//出//锤子//猛//击//张//的//后脑//，//将//张//击//倒//在//地//，//并//用//双手//掐//张//的//脖子//一//分//多//钟//，//张//昏//倒//过去//。//李//翻//箱//倒//柜//，//拿//走//１０００//元//现金//和//金//项链//、//金//戒指//、//金//耳坠//等//物//，//随后//将//张//的//尸体//拽//到//北//屋//，//从//橱//房//拿//出//一//桶//花生油//，//纵火//后//反//锁//门//逃走//。